

Web Analytics



AN INTRODUCTION
FOR EMORY UNIVERSITY
LIBRARIANS & STAFF

Some Examples

2

- Google Analytics Report
 - <http://www.google.com/analytics/>
- AWStats Report
 - http://www.nltechno.com/awstats/awstats.pl?month=04&year=2009&output=main&config=destailleur.fr&frame_name=index
- Webalyzer Report
 - <http://slavevoyages.org/usage>
- Robotic Crawlers Animated Report
 - <http://www.joanasmith.com/resources/googleCrawl.gif>

Goals

3

At the end of the course, the student should have a basic understanding of how website activity is monitored and analyzed.

You should be able to:

- ✦ Describe the kind of data typically captured by analytics tools
- ✦ Explain how analytics tools work
- ✦ Explain some of the limitations of such tools
- ✦ Explain some of the benefits of each type of tool
- ✦ Understand a Google analytics report (script-based)
- ✦ Understand a Webalyzer report (log-based)

Goal Summary:

- ✦ Have enough understanding of analytics to decide *what website data* would be useful for your job

Part 1

4

HOW THE WEB WORKS

Behind The Scenes

5

- **Web Servers**
 - Protocols
 - Applications
 - Logs
- **Content**
 - Mark-Up Languages
 - Databases
 - Images, Video, and more
- **Browsers**
 - Display
 - Persist
 - Remember

Telephones: a WWW Analogy

6

- **Telephones have many uses (=> Protocols)**
 - Voice calls
 - Fax
 - Dial-Up Internet
 - Automated Transactions (“Press 1 to place an order...”)
- **One number per telephone (=> IP Address)**
 - Many people on together (home extension handsets)
 - Many people at one number (business extension numbers)
 - Unknown caller (“unlisted number”)
- **Phone Number Look Up**
 - Yellow Pages (search engines)
 - White Pages (DNS)

The WWW & The Telephone Analogy

7

- The “Public Pay Phone” view of the world wide web
 - Only businesses have a phone number → a fixed IP address
 - People have to use public pay phones to reach the business
 - A URL becomes the business phone number you call
 - **The business does not call you!**
- DNS: The Web’s Phone Book
 - <http://larson.emory.edu/> = 170.140.223.107
 - DNS “looks up” the URL and then “calls” the IP address
- Calls come in from a Pay Phone
 - number = no definitive name
 - number = lots of names
- How to “name” that pay phone caller?

How The Web Works: Anatomy of a URL

8

- URLs come in 3 parts
 - Protocol
 - Site (Host)
 - Request
 - ✧ `http://www.foo.org/main.html`
 - ✧ `http://barfoo.net?verb=select&item=hat`
- Another explanation can be found here:
 - ✧ <http://xhtml.com/en/xhtml/media-types-how-the-web-works/>

Protocols

9

- HTTP
- HTTPS
- FTP
- SVN
- And lots more:

<http://www.realifewebdesigns.com/web-resources/web-protocols.html>

Like the “Voice”, “Fax”, and “Dial-up” protocols of telephones

Mark-Up Languages

10

- **HTML, XHTML, XML, VMRL, etc...**
 - Made browsing as we know it today possible
 - Browsers expanded web use -> uses
- **Multiple components**
 - Metadata (title, author, date)
 - Markup (tags/display instructions)
 - Content (data)
 - Locally-executed code (e.g., Javascript)
- **Variable presentation**
 - Internet Explorer vs Firefox
 - IE on Mac vs PC

Browsers

11

- “Presenters” for the web
- Read information
 - Decide what kind
 - ✦ Application
 - ✦ Mark-Up
 - ✦ Other
 - Act on decision
 - ✦ Launch application (Movie Player, PDF Reader)
 - ✦ Display Mark-Up (HTML/XHTML pages, text)
 - ✦ Present dialog (“What should Firefox do with this file?”)
- Store, Return, Record
 - Cookies (Remember who this is)
 - Click response (purchase this item; go to next page)
 - Data transfer (“user clicked here”)

Raw Web Exercise

12

- Downloading a web page the old-school way (the “telephone network” way)
 - ✧ <http://www.owenworks.biz/>
 - ✧ <http://www.simple.com/>
 - Browsers do this work for us automatically
- The page is more than just words, though
 - What happened to the images?
 - What’s all that other stuff?

Sample Page & Source

13

- <http://www.simple.com>
 - Metadata elements
 - Script to decide which browser
 - Links to images
- How does this compare with content received by the “old-school” method?

Key Concepts

14

- Page “Views”
 - What’s a page???
 - Hits
 - Sessions
 - Users
- Visiting simple.com and owenworks.biz:
 - Was this a hit?
 - A session?
 - A page view?
- Web analytics “guesses” at those answers
 - Logs, Scripts & Cookies provide hints

Quiz #1

15

1. What are the 3 parts of this URL:
 - <http://www.slavevoyages.org/tast/index.html>
2. Give an example of web page (HTML) metadata
3. Give an example of a mark-up tag
4. What is DNS?
5. What is a web hit?

Part 2

16

DATA CAPTURE METHODS

Cookies

17

- **Definition**
 - A piece of *text* stored on a PC
 - Sent to the PC by a website in the Header
 - Usually some name-value pair
 - PC sends it back to the website each time it returns to visit
 - ✦ Maintains “sessions”
 - ✦ Server script collects/watches the data (custom tool)
 - Examples in c:\windows\cookies
 - or in Safari->Preferences->Security->Show Cookies
- **It is *not*:**
 - Executable code
 - ✓ Let’s look at an Amazon cookie example
- **It can:**
 - Be controlled by you (refused, expired, expunged)
 - Make website more “user-friendly”
 - Help website track new vs repeat visitors
 - Watch repeat-visit frequency
 - Speed up purchasing (shopping cart tracking, quick checkout, etc.)
 - “Track” your movement through the web (Double-Click ad cookies, e.g.)

Web Page Scripts

18

- Operate *locally*
 - Your PC executes the code
 - Result **might** be sent to another machine
- Many uses
 - Calculations (convert Fahrenheit to Celsius)
 - Web page behavior tracking (Google Analytics)
 - Submit web form data to the server (Travel Expense Report)
- Code is *in the web page itself*
 - Example: page source from the slave voyages main page
 - Website scripts & Google Analytics code are both here

Sample Google Analytics Script

19

```
<!-- Google Analytics -->
```

```
<script type="text/javascript">
```

```
var gaJsHost = (("https:" == document.location.protocol) ? "https://ssl." :  
  "http://www.");
```

```
document.write(unescape("%3Cscript src='" + gaJsHost + "google-  
  analytics.com/ga.js' type='text/javascript'%3E%3C/script%3E"));
```

```
</script>
```

```
<script type="text/javascript">
```

```
var pageTracker = _gat._getTracker("UA-6500832-2");  
pageTracker._trackPageview();
```

```
</script>
```

```
<!-- Google Analytics -->
```

Web Server Logs

20

- Web servers track each request that comes in
 - One page may include many sub-requests:
 - ✦ Text Content (1 request)
 - ✦ Multiple Images (1 request per image)
 - Error Logs and Success Logs → usually combined
 - One log entry *per item*
- Common Log Format contains:
 - ✦ Where request came from (IP)
 - ✦ When (full time-stamp)
 - ✦ What (request)
 - ✦ Bytes sent (~file size)
 - ✦ Status code (success, failure, timeout, redirect, etc.)
 - ✦ Browser type
 - ✦ Referral source (search engine, e.g.)

Sample Log Data

21

1. Simple example:

```
66.249.71.140 - - [14/May/2009:00:25:42 -0400] "GET /projects/GarageFloorProject/garage4.jpg
HTTP/1.1" 304 - www.joanasmith.com "-" "Googlebot-Image/1.0" "-"
```

2. Search engine referral example:

```
146.151.104.125 - - [14/May/2009:00:03:18 -0400] "GET /acknowledgments.html HTTP/1.1" 200
7547 www.joanasmith.com "http://www.google.com/search?q=acknowledgments+in+a
+dissertation&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a"
"Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.5; en-US; rv:1.9.0.10) Gecko/2009042315
Firefox/3.0.10" "-"
```

FIELD ORDER = %h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\"

IP ADDRESS - - [TIMESTAMP] "Method Request Protocol" Status Code Bytes Host "Referer"
"User Agent" "-"

Note: %l = client id; %u=authenticated user id; not tracked here.

Also, "referer" is in fact misspelled in the original specification, and the tradition has "stuck"

Quiz #2

22

1. What is a Cookie?
2. What is a Script?
3. What's the difference between a web server log entry and a web page script?
4. What is a Referer (sic)?
5. List at least 4 items a web server log can track

Part 3

23

HITS, MISSES, & ERRORS IN DATA COLLECTION

Typical Data Collection Goals

24

- **Aggregating data into identifiable sessions**
 - Recognize user1 from user2
 - Track activity path for the session (while you're at the site)
- **Inferring user intent**
 - Why did the user come
- **Inferring user satisfaction**
 - Length of session
 - Number of pages visited
- **Determining “popular” pages**
- **Monitoring site problems**
- **And more...**

Data Collection Issues

25

- **Incomplete web page load**
 - Your analytics script might not run!
 - Where is it on the page?
- **Separating users from robots**
 - Robots crawl the web for information (Googlebot, Yahoo Slurp!, Msnbot)
 - Not all robots admit who they are
 - Very aggressive and thorough – use up lots of bandwidth
- **Logging problems and errors**
 - Full disk
 - Logging stopped
 - Logs over-written
 - Incomplete log record
- **Proxies**
 - Net Proxy (Network Address Translation) – users and sessions
 - Cache Proxy – page view counts
- **Internet “hiccups”**

Quiz #3

26

1. Why does a Net Proxy (NAT) make it hard to identify individual users?
2. Why does a Cache Proxy affect data collection?
3. What kind of data can be used to infer user satisfaction?
4. Cite one example of a web logging problem
5. Name one web robot

Web Analytics

27

END OF FIRST SESSION